Behav Genet (2013) 43:254–266
DOI 10.1007/s10519-013-9590-1

ORIGINAL RESEARCH

The Use of Imputed Sibling Genotypes in Sibship-Based Association Analysis: On Modeling Alternatives, Power and Model Misspecification

Camelia C. Minică · Conor V. Dolan · Jouke-Jan Hottenga · Gonneke Willemsen · Jacqueline M. Vink · Dorret I. Boomsma

Received: 12 September 2012 / Accepted: 18 February 2013 / Published online: 22 March 2013

Abstract When phenotypic, but no genotypic data are available for relatives of participants in genetic association studies, previous research has shown that family-based imputed genotypes can boost the statistical power when included in such studies. Here, using simulations, we compared the performance of two statistical approaches suitable to model imputed genotype data: the mixture approach, which involves the full distribution of the imputed genotypes and the dosage approach, where the mean of the conditional distribution features as the imputed genotype. Simulations were run by varying sibship size, size of the phenotypic correlations among siblings, imputation accuracy and minor allele frequency of the causal SNP. Furthermore, as imputing sibling data and extending the model to include sibships of size two or greater requires modeling the familial covariance matrix, we inquired whether model misspecification affects power. Finally, the results obtained via simulations were empirically verified in two datasets with continuous phenotype data (height) and with a dichotomous phenotype (smoking initiation). Across the settings considered, the mixture and the dosage approach are equally powerful and both produce unbiased parameter estimates. In addition, the likelihood-ratio test in the linear mixed model appears to be robust to the considered misspecification in the background covariance structure, given low to moderate phenotypic correlations among siblings. Empirical results show that the inclusion in association analysis of imputed sibling genotypes does not always result in larger test statistic. The actual test statistic may drop in value due to small effect sizes. That is, if the power benefit is small, that the change in distribution of the test statistic under the alternative is relatively small, the probability is greater of obtaining a smaller test statistic. As the genetic effects are typically hypothesized to be small, in practice, the decision on whether family-based imputation could be used as a means to increase power should be informed by prior power calculations and by the consideration of the background correlation.

Keywords Family-based imputation · Mixture model · Dosage model · Robustness

Introduction

Increasingly twin and family registries include both phenotypic data and genotypic data measured in family members (Boomsma et al. 2006; Willemsen et al. 2010). However, due to specific design or resources, the genotypic data may be limited to a subset of the family members, such as a single sibling. It is well recognized that limiting association analysis to 'the complete data participants', i.e., discarding relatives whose data are limited to phenotypic measures, is wasteful. As demonstrated by Visscher and Duffy (2006) and by Chen and Abecasis (2007) the genetic relations among the relatives can be used to impute genotypes of relatives lacking observed genotypic data. Subsequently including the relatives in the association study

Edited by Sarah Medland.

C. C. Minică (✉) · C. V. Dolan · J.-J. Hottenga ·
G. Willemsen · J. M. Vink · D. I. Boomsma
Department of Biological Psychology, VU University
Amsterdam, Van der Boechorststraat 1, 1081 BT,
Room 2B03, Amsterdam, The Netherlands
e-mail: c.c.minica@vu.nl

C. V. Dolan
Department of Psychology, University of Amsterdam,
Amsterdam, The Netherlands

will increase the power to detect association, although actual increase depends on the phenotypic correlations among the relatives (Visscher and Duffy 2006) and on the accuracy of the imputations (Chen and Abecasis 2007).

The goal of this article is to further investigate the factors affecting power following family-based imputation. We consider imputation of up to 3 sibling genotypes given a single genotyped sibling or a single genotyped sibling and one parent. Within these imputation setups we carry out an extensive comparison of the performance of the two statistical approaches, namely, the mixture model, which involves the full distribution of the imputed genotypes and the dosage approach, in which the mean of the conditional distribution features as the imputed genotype. The comparison is performed for two minor allele frequencies (MAF) and a range of background correlations. Sibling data only are included into the association analysis, where the sibhips vary from 1 (the genotyped sib) to 4 (1 observed, 3 imputed genotypes). To check the validity of our simulation program and the power calculations, we also report the power in the full information model, as an indication of the maximum power, attainable when all siblings in a sibship have observed genotypes.

Secondly, we examined the effect on power of misspecification of the background covariance structure in family-based association analysis. Imputing genotypes and extending the model to include sibships of size two and greater does require modeling the background covariance matrix. Such modeling may be of interest substantively, or as a means to reduce the parameter space. As the calculation of power to detect a measured (imputed) genetic effect will require some choice of background covariance structure, one may ask whether misspecification will affect the statistical power. To address this question, we simulate sibling phenotypes according to an additive genes/unique environment (AE) model and next, we fit two alternative models to these data: a correctly specified AE model, consistent with the model used for simulation, and a misspecified common environment/unique environment (CE) model. We compare the observed powers of the two models, with and without the misspecification. As model misspecification is of interest regardless of whether genotypes are imputed or not, we study its effect on power both in the 'all genotypes observed' setting (i.e., the full information setting) and in the setting in which some genotypes were imputed (i.e., the dosage setting).

Finally, we illustrate empirically the results obtained using simulations. In one empirical dataset we sought to quantify the power gains conferred by family-based imputation when the trait of interest is assessed on a continuous scale. This analysis aims to replicate 112 of the 180 height single nucleotide polymorphisms (SNPs) reported by Lango Allen et al. (2010) in a Netherlands Twin Register (NTR) dataset consisting of 5,910 siblings with observed and imputed genotypes. We explore the mixed results by means of the analysis of simulated data. The second illustration considers tests of association between observed and kinship-based imputed SNPs and a discrete trait—smoking initiation. Specifically, in a dataset comprising of 5,981 observed and imputed sibling genotypes we reran the analysis of Vink et al. (2009) for 20 of the 41 SNPs associated with smoking initiation in their discovery sample. Both analyses used solely sibling data and were carried out first in the 'complete data' samples and then by extending the samples to include the imputed sibling genotypes.

Methods

Models for sibship-based association

We simulated genotypic and phenotypic data for nuclear families with four siblings. In the full information setting, we computed the power to detect genetic association using the complete information, i.e., 1 to 4 sibling genotypes and phenotypes. Next, we limited the genotypic information to 1 sibling, or to 1 sibling and 1 parent, and, conditional on this information we calculated the missing genotype distribution in the remaining siblings. In this limited information setting we considered the power of the mixture model and of the dosage approach. Below we provide the details of the three modeling approaches and of our simulations.

The full information model

We considered a diallelic locus with alleles $\mathbf{A}$ and $\mathbf{a}$, and frequencies $\mathbf{p}$ (A) and $\mathbf{q} = 1 - \mathbf{p}$ (a), observed in nuclear families with four siblings. Let $\mathbf{g}_\mathrm{i}$ denote the vector of genotypes of $\mathbf{m}$ (1 to 4) sibs in family $\mathbf{i}$, where possible elements of $\mathbf{g}_\mathrm{i}$ are $\mathbf{AA}$, $\mathbf{Aa}$, and $\mathbf{aa}$ (Falconer and Mackay 1996). Throughout, the locus has an additive effect on the phenotype, so we can assign the values $\mathbf{d}$, $\mathbf{0}$ and $-\mathbf{d}$ to the three possible genotypes, where the value of $\mathbf{d}$ is dictated by the minor allele frequency and our effect size. Letting the allele $\mathbf{A}$ be increaser allele, we code $\mathbf{1}$ for the genotype $\mathbf{AA}$, $\mathbf{0}$ for the genotype $\mathbf{Aa}$, and $-\mathbf{1}$ for the genotype $\mathbf{aa}$. Let $\mathbf{x}_\mathrm{ij}$ denote the vector of genotype indicators ($-\mathbf{1}$, $\mathbf{0}$ or $\mathbf{1}$). We regressed the phenotypes $\mathbf{y}_\mathrm{i}$ observed in $\mathbf{m}$ sibs in family $\mathbf{i}$ (i.e., $\mathbf{y}_\mathrm{i}^\mathrm{t} = [\mathbf{y}_{\mathrm{i}1} \ldots \mathbf{y}_{\mathrm{im}}]$, where $\mathbf{t}$ denotes transposition) on the indicators:

$$\mathbf{y}_\mathrm{i} = \mathbf{b}_0 + \mathbf{x}_\mathrm{i} * \mathbf{b}_1 + \mathbf{e}_\mathrm{i}, \qquad (1)$$

where $\mathbf{b}_0$ is an $\mathbf{m}$ vector containing the intercept (e.g., for $\mathbf{m} = 4$, the elements of $\mathbf{b}_0$ are $\mathbf{b}_0^\mathrm{t} = [\mathbf{b}_0\ \mathbf{b}_0\ \mathbf{b}_0\ \mathbf{b}_0]$), $\mathbf{b}_1$ is the scalar parameter of main interest, and $\mathbf{e}_\mathrm{i}$ is the $\mathbf{m}$ vector of residuals. Conditional on genotype, the means are

$\mathbf{\mu}_1 = \mathbf{b}_0 + \mathbf{b}_1$, $\mathbf{\mu}_2 = \mathbf{b}_0$, or $\mathbf{\mu}_3 = \mathbf{b}_0 - \mathbf{b}_1$, and the residuals are distributed $\mathbf{e}|\mathbf{x} \sim \mathbf{N}(0,\mathbf{S}_0)$, where $\mathbf{S}_0$ is the $\mathbf{m} \times \mathbf{m}$ positive definite covariance matrix. In the OpenMx specification, the background covariance matrix was estimated using the decomposition $\mathbf{S}_0 = \mathbf{DD}^t$, where $\mathbf{D}$ is an unconstrained lower triangular matrix.

We refer to this model as the full information model, as this model is based on the complete genotype information measured in all siblings in the sibship, i.e., all elements of $\mathbf{x}_i$ are observed. In this setting, the power analyses were based on both exact data simulations (Van der Sluis et al. 2008) and on the standard Monte Carlo procedure. In the latter, power was computed as the proportion of analyses in which minus twice the difference in the log likelihoods the two models—with and without the genotypic effect—is greater than a critical value associated with the chosen alpha (i.e., $\mathbf{c}_\alpha = 6.64$ given $\mathbf{\alpha} = .01$). The Monte Carlo procedure was employed for consistency: in the mixture approach, we do not have sufficient statistics and therefore cannot conduct exact power calculations.

The mixture approach

We considered the situation in which phenotypic data have been collected in sibships of sizes 2, 3, and 4, while genotypic data are limited to 1 sibling, or to 1 sibling and 1 parent.

Conditional on sib 1 genotype ($\mathbf{g}_{i1}$), we calculated the probability of the sibling $\mathbf{j}$ ($\mathbf{j} = 2...\mathbf{m}$) genotype ($\mathbf{g}_{ij}$) as

$$\text{prob}(\mathbf{g}_{ij}|\mathbf{g}_{i1}) = \text{prob}(\mathbf{g}_{ij}\& \ \mathbf{g}_{i1})/\ \text{prob}(\mathbf{g}_{i1}) \tag{2}$$

(Chen and Abecasis 2007). The probabilities prob($\mathbf{g}_{ij}$ & $\mathbf{g}_{i1}$) and prob($\mathbf{g}_{i1}$) can be derived from Mather and Jinks (1977, ch. 7). Given $\mathbf{m}$ sibs, we calculate 3^{m-1} conditional probabilities given the sib 1 genotype. This procedure is followed for size 3 and 4 sibships, where conditionally on the genotypic information within a family, the siblings 2 to 4 genotypes are independent events.

Equation 2 can be extended to include parental genotype ($\mathbf{g}_p$) if this is available additionally to the sib 1 genotype. Thus, more accurate conditional probabilities of the sib $\mathbf{j}$ ($\mathbf{j} = 2...\mathbf{m}$) genotype are obtained as:

$$\text{prob}(\mathbf{g}_{ij}|\mathbf{g}_{i1}\& \ \mathbf{g}_p) = \text{prob}(\mathbf{g}_{ij}\& \ \mathbf{g}_{i1}\& \ \mathbf{g}_p)/\ \text{prob}(\mathbf{g}_{i1}\& \ \mathbf{g}_p) \tag{3}$$

Again the relevant probabilities can be derived from Mather and Jinks (1977). To provide an indication of the values of the posterior probabilities, these are shown in Table 1 for MAF of .2. Table 1 includes the unconditional Hardy–Weinberg (H–W) probabilities and the genetic probability index (GPI; Kinghorn 1997), which is a measure of the distance of the imputed probabilities to the H–W probabilities. The measure ranges from 0 (H–W probabilities) to 100 (genotype observed). We return to this measure in the discussion.

Table 1 Posterior probabilities of the sibling 2 (s2) genotype AA, Aa, or aa, conditional on the observed genotype in a single sib (s1) or in a single sib and a single parent (p1), and given MAF = .2. The H–W probabilities are the unconditional probabilities. The GPI is Kinghorn's genetic probability index, a distance measure (ranging from 0 to 100) of the imputed probabilities from the H–W probabilities

Observed	Posterior probabilities of the s2 genotype			GPI
	AA	Aa	aa	
None (H–W)	.04	.32	.64	0
s1 AA	.36	.48	.16	49.33
s1 Aa	.06	.58	.36	38.67
s1 aa	.01	.18	.81	47.29
s1 AA and p1 AA	.60	.40	.00	68.92
s1 Aa and p1 AA	.10	.90	.00	86.67
s1 AA and p1 Aa	.30	.50	.20	45.33
s1 Aa and p1 Aa	.10	.50	.40	28.65
s1 aa and p1 Aa	.05	.50	.45	26.69
s1 Aa and p1 aa	.00	.60	.40	41.38
s1 aa and p1 aa	.00	.10	.90	72.27

For instance, given **aa** observed in sib 1, the genotype probabilities of **AA**, **Aa**, and **aa** are .01, .18, and .81, respectively. Given **aa** observed in sib 1 and in the parent, these probabilities are .0, .10, and .90.

To test for association, we fitted a mixture model that incorporates the regression model defined in Eq. (1). That is, we regressed the observed phenotypes on the possible elements of $\mathbf{x}_{ij}$ (i.e., 1, 0, −1), and we weighted the associated densities by the conditional probabilities calculated conditional on sib 1, or on sib 1 and parent 1.

The mixture fitted to the data is a $3^{\mathbf{m}-1}$ component mixture, where the proportion of sibpair genotypes within each component of the mixture is determined by the conditional probabilities (i.e., the finite mixture proportions). For example, consider a sibship of size 2, where we have at our disposal the phenotypes observed in both siblings $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$, the genotype observed in sib 1 ($\mathbf{g}_{i1}$) and 3 probabilities based on $\mathbf{g}_{i1}$ (and on parental genotype $\mathbf{g}_p$, if available). Conditional on the sib 1 observed genotype (and possibly $\mathbf{g}_p$) the distribution of the vector $\mathbf{y}_i$ of the observed phenotypes is assumed to follow a three component bivariate normal mixture. This mixture distribution can be expressed as the sum of 3 component distributions weighted by the fixed mixing proportions $\mathbf{p}_k$ (i.e., the probabilities, conditional on the observed genotype, of impute genotype **AA**, **Aa**, or **aa**) of sib-pairs in each component:

$$\mathbf{f}(\mathbf{y}_i; \mathbf{p}, \mathbf{S}, \boldsymbol{\mu}) = \Sigma_{k=1}^{3} p_k N_k(\mathbf{y}_i; \mathbf{S}_0, \boldsymbol{\mu}_k) \quad (4)$$

in which $\mathbf{S}$ equals $[\mathbf{S}_0\ \mathbf{S}_0\ \mathbf{S}_0]$, the matrix $\boldsymbol{\mu}$ contains the means vectors of each component $[\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3]$ (possible elements of $\boldsymbol{\mu}_k$ are $\mathbf{b}_0 + \mathbf{b}_1$, $\mathbf{b}_0$, and $\mathbf{b}_0 - \mathbf{b}_1$), $\mathbf{p} = [\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3]$ where $\mathbf{p}_k$ represents the fixed mixing proportions of sib-pairs within the $\mathbf{k}$-component distribution, and $N_k(\mathbf{y}_i; \mathbf{S}_0, \boldsymbol{\mu}_k)$ represents the k-variate normal density function within each component. The number of components conditional on the sib 1 genotype is $3^{\mathbf{m}-1}$, hence in the case of 3 (4) siblings, we have 9 (27) components. As in the full information setting, in the specification in OpenMx, we modeled the background covariance matrix using the decomposition $\mathbf{S}_0 = \mathbf{D}\mathbf{D}^t$. We imposed no additional constraints on $\mathbf{D}$.

The dosage approach

In this approach, we calculated the expected value of the genotype indicator based on the conditional probabilities estimated as defined by the Eqs. (2) or (3). That is, conditional on the sib 1 genotype ($\mathbf{g}_{i1}$), and given our coding of 1 (**AA**), 0 (**Aa**), and −1 (**aa**), the average indicator is calculated as:

$$\mathbf{x}_{ij}^{\bullet} = \text{prob}(\mathbf{g}_{ij} = \text{AA}|\mathbf{g}_{i1}) - \text{prob}(\mathbf{g}_{ij} = \text{aa}|\mathbf{g}_{i1}) \quad (5)$$

where $\mathbf{x}_{ij}^{\bullet}$ represents the vector of expected number of increaser alleles in sib $\mathbf{j}$, and $\mathbf{x}_i^{\bullet} = [x_{i1}\ x_{i2}^{\bullet}...x_{im}^{\bullet}]$ in family $\mathbf{i}$. We specify the regression model for the observed phenotype $\mathbf{y}$ in family $\mathbf{i}$ as follows:

$$\mathbf{y}_i = \mathbf{b}_0 + \mathbf{x}_i^{\bullet} * \mathbf{b}_1 + \mathbf{e}_i \quad (6)$$

where, as above, $\mathbf{b}_0$ is an $\mathbf{m}$ vector containing the intercept, $\mathbf{b}_1$ is a scalar parameter, and $\mathbf{e}_i$ is the $\mathbf{m}$ vector of residuals. The residuals are distributed approximately as $\mathbf{e}|\mathbf{x} \sim \mathbf{N}(\mathbf{0},\mathbf{S}_1)$. The subscript serves to indicate that the conditional covariance matrices—$\mathbf{S}_0$ and $\mathbf{S}_1$—are not expected to be exactly equal, as the variance of $\mathbf{e}_{ij}|\mathbf{x}_{ij}^{\bullet}$ is slightly lower than the variance of $\mathbf{e}_{ij}|\mathbf{x}_{ij}$ (Visscher and Duffy 2006; Chen and Abecasis 2007). As in the previous approaches, the expected background covariance matrix is modeled using the Cholesky decomposition. We computed the power to detect genetic association in this model by the means of the Monte Carlo procedure.

Model fitting

We implemented the three models in OpenMx (R package version 1.0.5; Boker et al. 2011). The full information model and the dosage model were also implemented in the R-nlme package (using the lme function; Pinheiro et al. 2012). This implementation is identical to the OpenMx implementation, except that the conditional covariance matrix was constrained as $\mathbf{S_1} = \mathbf{J}\sigma_A^2\mathbf{J}^t + \sigma_e^2\mathbf{I}$, where $\mathbf{J}$ is the $\mathbf{m}$ x $\mathbf{1}$ unit vector, and $\mathbf{I}$ is the ($\mathbf{m}$ x $\mathbf{m}$) identity matrix. This specification is consistent with the simulation in the full information model, but slightly misspecified in the dosage model: as mentioned above, the variance of $\mathbf{e}_{ij}|\mathbf{x}_{ij}^{\bullet}$ ($\mathbf{j} = 2...\mathbf{m}$) is lower than the variance of $\mathbf{e}_{ij}|\mathbf{x}_{ij}$. We expect this misspecification to be trivial, as the effect size of the QTL is small (1 %; see below). In all cases the models were fitted by means of maximum likelihood estimation.

Simulation details

1,000 genotypic and phenotypic datasets comprising 500 nuclear families with 4 siblings were simulated in R (R development core team 2005). We first simulated parental genotypes at a single diallelic locus in H–W equilibrium and, given random mating, we used these to generate the sibling genotypes. We assumed the diallelic genotype explained 1 % of the phenotypic variance. As mentioned above, we varied the minor allele frequencies (.2 and .5) and the background phenotypic correlations among siblings (.2 to .8, by 2). Note that the effect size was 1 % regardless of MAF. We calculate and report the increase in power relative to an association analysis which includes only the subjects with observed genotypes, given the α of .01. All simulations were carried out using the R software package (http://www.r-project.org/) and were run on the Genetic Cluster Computer (http://www.geneticcluster.org).

Misspecification of the background covariance structure

Next, we studied the effect of misspecification of the background covariance matrix (i.e., more serious than the difference between $\mathbf{S}_0$ and $\mathbf{S}_1$) in family-based association analysis. Sibling phenotypes were simulated according to an AE model, which included: (a) a SNP with equally frequent alleles, accounting for 1 % of the phenotypic variance, (b) background heritability of .8, .45 and .15, and (c) unshared environmental effects. We considered nuclear families with sibship size 2 [pairs of monozygotic (MZ) and dizygotic (DZ) twins] and 4 (pairs of twins and 2 siblings), where genotypes as well as phenotypes were observed in all siblings in the sibship (the full information setting). Furthermore, we simulated the limited information setting, where some genotypes were missing (the dosage setting). That is, in this latter setting, 50 % genotypes were missing among parent 1 and parent 2 and 50 % genotypes were missing among each sibling in the sibship.

To model association, two alternative models were fitted to the AE simulated data: (a) the correctly specified AE model, and (b) a CE model, where the background correlations among siblings in the sibship were (incorrectly) constrained to be equal. To obtain empirical estimates of the power, we carried out 10,000 replications and we computed the proportion of datasets in which the genetic effect was detected, given four levels of significance ($\alpha = 10^{-2}$, 10^{-3}, 10^{-4}, and 10^{-7}).

In addition, we verified the type I error rates, in both settings, when fitting the model with and without the misspecification. For this, sibling data were simulated under the null model of no association given the conditions described above; we then evaluated the effect of background misspecification on the type I error rates at alpha levels of 10^{-2}, 10^{-3}, and 10^{-4} by examining 10,000 replicates (100,000 replicates for the $\alpha = 10^{-4}$ cell).

Empirical illustrations

Height data

We illustrated empirically the results obtained using simulations. The first analysis examines the power advantages conferred by family-based imputation when the trait tested for association is continuous. First we performed family-based imputation of 112 of the 180 SNPs previously associated with height (Lango Allen et al. 2010) and next, we carried out a sibship-based association analysis. We ran the analysis with and without the imputed sibling genotypes, and we assessed the association signals in the two samples.

The data set used for this illustration consisted of 2,164 Dutch nuclear families from the NTR, where observed or self-reported height data were available for 5,910 siblings born between 1914 and 1991 ($N = 3{,}667$ females with a mean height of 169.89 cm and SD = 6.43 cm, and $N = 2{,}243$ males with a mean height of 183.16 cm and SD = 7.07 cm). Families were included if at least one member had observed genotypes. Height was measured in adults at 18 years or older, and data of individuals with multiple measurements available underwent consistency checks (i.e., 236 siblings, representing 1.3 % of the 17,195 siblings who formed the initial phenotypic sample were discarded due to differences larger than 5 cm between multiple measures). As imputation exploits biological relationships within nuclear families, we also excluded self-reported half-siblings and non-biological parents ($N = 108$ individuals, .5 % of the phenotypic sample). Genotypic data were limited to 2,410 siblings and 1,437 parents. Conditional on the observed genotypes we imputed 3,500 siblings who had height but no genotype data. To impute missing sibling genotypes we used our own R script (CSIBPROB, see http://www.psy.vu.nl/nl/over-de-faculteit/medewerkers-alfabetisch/medewerkers-mp/minica-c-c/index.asp).

In the next step, we carried out a linear mixed association analysis (Visscher et al. 2004), first by limiting the sample to the observed genotypes and second, by extending the sample to incorporate imputed siblings genotype data by using genotype dosages. Height was regressed on the genotype indicator variable and on the observed covariates (sex and birth cohort) modeled as fixed effects. As the sample included monozygotic twins (i.e., $N = 656$ MZ twin pairs) and full siblings, we modeled the background covariance structure by an AE model. Like in the simulations, the association analysis was limited to the sibling data.

The analysis of smoking initiation

The second empirical example illustrates the power gains obtained by the inclusion into an association analysis of imputed sibling genotypes when the phenotype of interest is dichotomous. Specifically, we reran the association analysis conducted by Vink et al. (2009) for 20[1] SNPs of the 41 SNPs associated with smoking initiation (at p-values $< 10^{-4}$) in their discovery sample. The original analysis was ran in unrelated individuals ($N = 3{,}497$), while the present one is sibhip-based, performed by implementing the above described two-step approach (i.e., imputation of missing sibling genotypes, which are subsequently incorporated in an association analysis).

Measured phenotypes were available for 17,641 siblings in 10,200 Dutch nuclear families from the NTR. Based on self-report, half-siblings ($N = 78$) and non-biological parents ($N = 192$) were excluded (representing .9 % of the initial phenotypic sample). As in the previous empirical example, solely families with at least one parental or sibling observed genotype were retained for the analysis. There were 2,210 families that met this criterion. In these families 2,458 siblings and 1,420 parents had observed genotypic data which were exploited to impute siblings with measured phenotypes but lacking genotypic data. The final phenotypic sample comprised of 3,125 controls (never smoked tobacco) and 2,856 cases (ever smoked tobacco); the siblings were born between 1914 and 1993 (mean age = 42.62 years, SD = 11.61) and 61 % of the sample were females. There were 86 siblings with observed genotypes but no smoking-initiation data.

To model association we used an AE generalized mixed effects model, fitted firstly to the sample limited to the 'complete data' siblings, and then to the sample incorporating siblings with imputed genotypes by using dosages. Sex and age were included as covariates. Model fitting was performed by using the MASS package (the function glmmPQL, Venables and Ripley 2002) and the nlme package for R (Pinheiro et al. 2012).

[1] Of the 41 SNPs, 20 SNPs were available in the current sample.

Results

The full information setting

We first evaluated the power to detect association in the full information setting to obtain an indication of the maximum power given maximum information (i.e., all siblings in the sibship have measured genotypes and phenotypes). This verifies the validity of our simulation program and our subsequent power calculations.

The results are displayed in Fig. 1 for the exact calculations and numerical values are shown in Table 2. As mentioned, the effect size was chosen to equal 1 % regardless of MAF, so that these results apply equally to MAF = .2 and MAF = .5.

Figure 1 (left) demonstrates the effect of the background correlation on power, in 500 families comprising size 1, 2, 3, or 4 sibship. The differences in power between the sibships sizes are expected given the differences in sample sizes (500 singletons confer less power than do 500 size 4 sibships). This is of little concern as we are interested in the change in power associated with the use of imputed genotypes within each sibship size. However, merely for comparison, we also calculated the power for a constant number of individual cases, specifically, 125 size 4 sibships, 166 size 3 sibships, 250 size 2 sibships, and 500 singletons. These results are shown in Fig. 1 (right). As Visscher et al. (2008) noted, power suffers when related individuals are included into analysis for small to moderate phenotypic correlations. However, for larger phenotypic correlations, the power of a family based design exceeds the power of an association analysis conducted in unrelated individuals, given constant genotyping resources.

The mixture and dosage approaches

Next, we considered the genotypic sample consisting of both observed and imputed sibling genotypes, and within this setting we examined the power and the estimation precision of the mixture model and of the dosage approach. Figure 2 depicts the results of the power analyses. We plotted the power relative to the expected power afforded by a sample size of 500 singletons, given the alpha of .01. The actual power in this case is .37 (Table 2), but this is scaled to equal 1, and the observed power is divided by this .37.

Table 2 Power in the full information model given an effect size of 1 %, α = .01 and $N = 500$ families. Power is shown as a function of the sibship size (nsib) and background correlation. In the case of a singleton (nsib = 1), the background correlation is not relevant

	background correlation			
nsib	.2	.4	.6	.8
4	.93	.95	.98	.99
3	.85	.86	.93	.99
2	.68	.69	.76	.93
1	.37	.37	.37	.37

Across all settings considered here, there was no difference in the observed powers of the mixture model and the dosage approach. We found the power of the two approaches was similarly affected by three factors: the phenotypic correlation (see also Visscher and Duffy 2006), the sibship size (2 to 4), and the accuracy of the imputation (based on 1 sibling or on 1 sibling and 1 parent).

When the imputation was based on 1 genotyped sibling, appreciable increase in power is observed only given relatively strong or weak background phenotypic correlations among the sibs. That is, when the background correlations were either small (i.e., <.4) or high (i.e., >.6) imputing siblings increased power by about a factor of 1.2–2 relative to ‘no imputation analysis’. Phenotypic correlations had a similar, albeit weaker effect, on the power given imputation based on 1 sibling and 1 parent genotypes. Within this setting, the association analysis including imputed sibling genotypes had greater power given low and high phenotypic correlations and it had reduced power for moderate phenotypic correlations. However, even for phenotypic

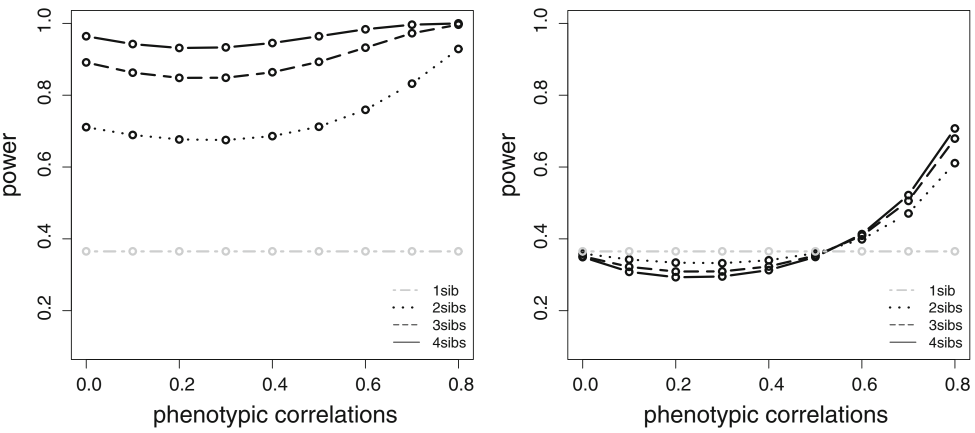

Fig. 1 The expected power in the full information setting for various background correlations, given α = .01, MAF = .2 and an effect size of 1 %. *Left* 500 families with 1, 2, 3 and 4 siblings. *Right* 500 genotyped siblings regardless of sibship size (i.e., 500 singletons, 250 size 2 sibships, 166 size 3 sibships, and 125 size 4 sibships)

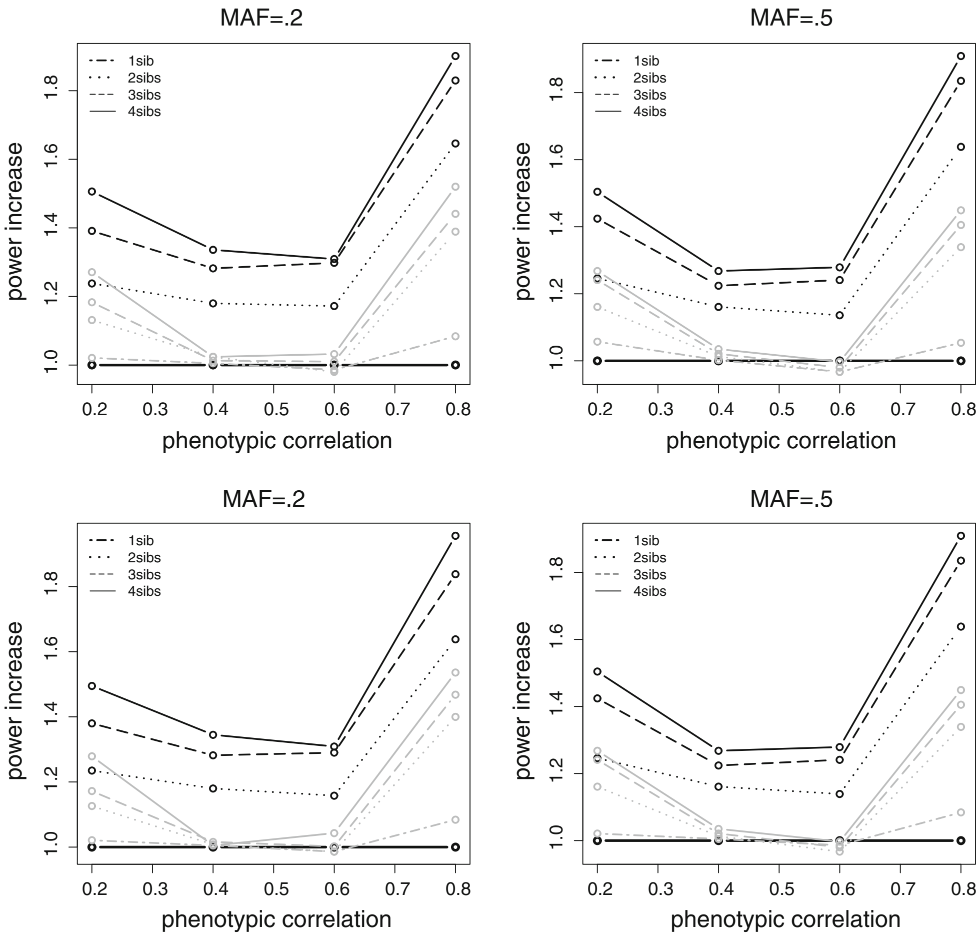

Fig. 2 The empirical power of the dosage model (*top*) and the mixture model (*bottom*), relative to the expected power afforded by 500 singletons (*the black bolded line*), given α = .01. *The grey lines* the empirical power afforded by sibships sizes 2, 3 and 4 when imputation is based on 1 genotyped sibling. *The black lines* the empirical power afforded by sibships sizes 2, 3 and 4 when imputation is based on 1 sibling and 1 parental genotypes. Power calculations are based on 1,000 datasets comprising 500 families, each dataset with a simulated genetic variant explaining 1 % of the phenotypic variance, regardless of MAF

correlations in this range this analysis was about a factor of 1.2–1.3 more powerful than the 'no imputation' analysis.

The power also increased with increasing sibship size. Apart from the moderate correlations condition, power was always larger in larger sibships, where, for instance, a size 4 sibship was about 10 % more powerful than a size 2 sibship.

Furthermore, as is to be expected, the design in which imputation was based on 1 parent and 1 sibling genotypes was found consistently more powerful than the design in which the imputation was based on 1 sibling genotype only, with average power gains of 10–15 %, across all conditions. In this setting, the additional information about parental genotype allowed an increase in the accuracy of the imputed genotypes, an increase that resulted in greater precision of estimating the genetic effect, and therefore was associated with greater power.

Tables 3 and 4 display the mean and the standard deviation of the estimate of $\mathbf{b}_1$ for MAF = .2 obtained in the mixture model and in the dosage model, as fitted in OpenMx (MAF = .5 produced comparable results). The averages of the estimate of the genetic effect $\mathbf{b}_1$ are close to their true value both when the analysis is limited to the observed genotypes and when it additionally includes imputed siblings. The variation in the standard deviation of

Table 3 Average estimates of the genetic effect $\mathbf{b}_1$ and the associated standard deviations (in parenthesis) for the mixture models, for MAF = .2. The true parameter value is $\mathbf{b}_1$ = .1767 (1,000 replicates)

Models	Sibship size	Background correlations			
		.2	.4	.6	.8
Observed genotypes	1	.176 (.079)	.176 (.077)	.175 (.078)	.180 (.076)
Conditional probabilities given 1 sibling genotype ($\mathbf{g}_{i1}$)	2	.176 (.075)	.176 (.077)	.175 (.078)	.180 (.070)
	3	.176 (.073)	.176 (.076)	.175 (.078)	.180 (.067)
	4	.177 (.073)	.176 (.076)	.176 (.077)	.180 (.064)
Conditional probabilities given 1 sibling and 1 parent genotypes ($\mathbf{g}_{i1}$&$\mathbf{g}_p$)	2	.176 (.070)	.176 (.073)	.175 (.071)	.178 (.062)
	3	.176 (.067)	.176 (.070)	.176 (.067)	.177 (.057)
	4	.176 (.064)	.175 (.068)	.176 (.066)	.177 (.053)

Table 4 Average estimates of the genetic effect $\mathbf{b}_1$ and the associated standard deviations (in parenthesis) for the dosage models, for MAF = .2. The true parameter value is $\mathbf{b}_1 = .1767$ (1,000 replicates)

Models	Sibship size	Background correlations			
		.2	.4	.6	.8
Observed genotypes	1	.176 (.079)	.176 (.077)	.175 (.078)	.180 (.076)
Dosage conditional on 1 sibling genotype	2	.176 (.075)	.176 (.077)	.175 (.078)	.180 (.070)
	3	.176 (.074)	.176 (.077)	.175 (.078)	.180 (.067)
	4	.177 (.073)	.177 (.076)	.176 (.077)	.181 (.066)
Dosage conditional on 1 sibling and 1 parent genotypes	2	.176 (.070)	.176 (.073)	.176 (.071)	.179 (.063)
	3	.176 (.067)	.176 (.071)	.176 (.067)	.178 (.059)
	4	.176 (.065)	.176 (.069)	.177 (.067)	.178 (.056)

the parameter estimate reflects the variation in power. Including siblings with missing genotypes yields unbiased estimates of the genetic effect and, as it leads to an increase in the sample size, it allows for higher estimation accuracy.

The results obtained using the dosage model implemented in OpenMx and nlme are quite similar (results not shown), notwithstanding that the background covariance matrix is highly constrained in nlme, but unconstrained in OpenMx.[2] This is expected as in the nlme specification the model for the background covariance matrix is almost completely consistent with the data generating model (the minor difference stemming from the differences between $\mathbf{S}_0$ and $\mathbf{S}_1$, as mentioned above).

The effects on power of misspecification of the background covariance structure

Figure 3 displays the results for the full information setting.

Figure 3 (left) indicates that in the full information setting the observed power of the misspecified CE model was in good agreement with the power of the correctly specified AE model for weak to moderate background correlations. With an increase in the background correlations we noted a slight discrepancy among the powers of the two models (Fig. 3, right). The discrepancy is higher (up to about 9 %) for the size 2 sibship than for the size 4 sibship. Results for the dosage model were similar (data not shown).

We also assessed the empirical type I error rates. Results for both the full information setting and the dosage model are given in Table 5.

As can be seen in Table 5 results were akin in the two settings: they indicate that for low and moderate background correlations, the misspecification of the background covariance structure yields empirical type I error rates that are consistent with the specified alpha levels. With these settings, the likelihood-ratio test in the linear mixed model appears to be robust to the degree of misspecification of the family structure considered here. However, one can note that when the background correlations are high (MZ correlation = .80), in the incorrect model the rate of type I errors is higher than expected. This effect is stronger in the size 2 sibship than in the size 4 sibship where the misspecification pertains to a single element of a 4 × 4 covariance matrix. Finally, we note that given the scenarios considered, the full information model and the dosage approach yielded similar results, confirming that imputation per se does not affect the type I error rates (see also Chen and Abecasis 2007).

Application: height data

The results of the sibship-based association analysis aimed at replicating 112 height SNPs in the NTR sample are illustrated in Fig. 4.

Imputation enhanced the association signal at some loci, notwithstanding that the sibling correlations are in the region where the power gains are lowest (i.e., siblings are correlated about .45 for height, e.g., Visscher et al. 2007). To provide an illustration, we show in Table 6 the markers—associations with p-values $<10^{-2}$ based on the observed data—for which we obtained the largest increase in χ^2 by including into analysis imputed sibling genotypes.

One SNP only—rs1351394—reached a significant association with height (p-value < .01/112), and clearly the association signal was stronger in the sample that included imputed siblings genotypes, i.e., $\chi^2 = 20.599$ versus $\chi^2 = 19.711$ in the no imputation analysis, respectively. In addition, we report the associations with a p-value < .01, as the present sample comprising 5,910 observed and imputed sibling genotypes was underpowered to yield more significant Bonferroni[3] corrected results. These results indicate that imputation increased the power to detect association, which

[2] Fitting the constrained model in Mx and nlme produced identical results.

[3] For convenience we have chosen the Bonferroni method to correct for multiple testing, although this procedure can be conservative (Laird and Lange 2011). However, in Fig. 4 we plot the values of the noncentrality parameter of the likelihood ratio test, as these values do not depend on the chosen alpha, or the correction for multiple testing. They are illustrative of the variation in power—before and following imputation—given various effect sizes.

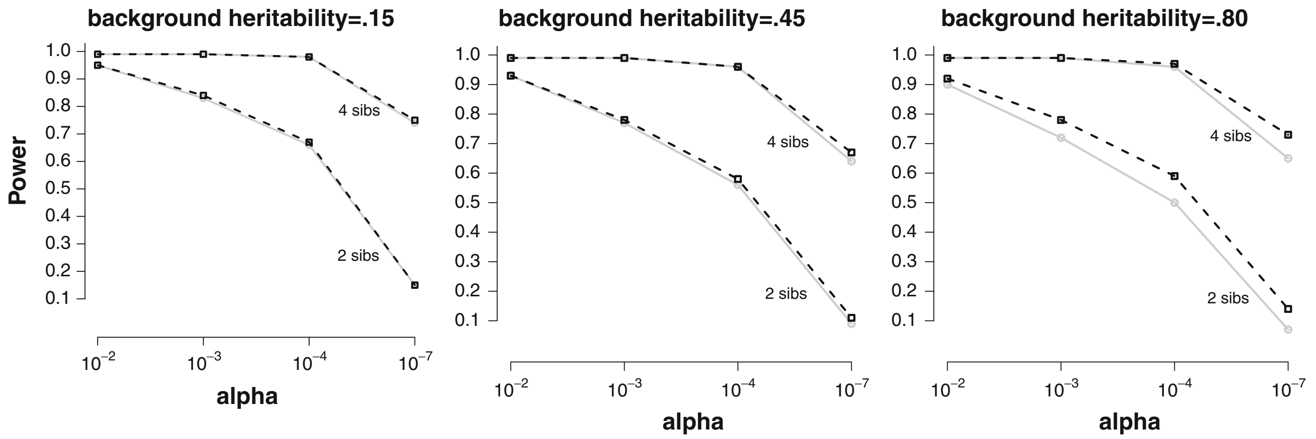

Fig. 3 The empirical power to detect a genetic variant with a MAF = .5, that explains 1 % of the trait variance in the correctly specified AE linear mixed model (*the grey line*) and in the misspecified CE linear mixed model (*the black dashed line*). In the correct model the background covariances among identical twins were specified as twice larger than in fraternal twins. In the incorrect model the background covariance matrix was estimated subject to equal covariances. The empirical power was computed for 10,000 datasets (100,000 datasets for the 10^{-7} cell) consisting of 500 MZ and 500 DZ families with sibships of size 2 and 4

Table 5 Type I error rates in the Full information and in the Dosage settings, in the correctly specified model (AE background) and in the misspecified model (CE background, results displayed in italics). We simulated sibling phenotypes for 500 monozygotic and 500 dizygotic families and a SNP having a MAF = .5 and explaining 1 % of the phenotypic variance. We varied the sibship size and the magnitude of the MZ background correlations (10,000 simulations/cell for the cells $\alpha = 10^{-2}$ and $\alpha = 10^{-3}$; 100,000 replicates for the $\alpha = 10^{-4}$ cell)

Sibship size	Background correlations	Level of significance	No missing genotypes AE/CE	Observed and imputed genotypes AE/CE
2	.15	$\alpha = 10^{-2}$	.010/*.010*	.009/*.009*
		$\alpha = 10^{-3}$	.001/*.001*	.001/*.0009*
		$\alpha = 10^{-4}$	.0001/*.0001*	.00007/*.00008*
	.45	$\alpha = 10^{-2}$	.010/*.012*	.010/*.012*
		$\alpha = 10^{-3}$	.001/*.001*	.0008/*.001*
		$\alpha = 10^{-4}$	.00007/*.0001*	.00009/*.0001*
	.80	$\alpha = 10^{-2}$	.009/*.01*	.01/*.01*
		$\alpha = 10^{-3}$	.001/*.002*	.001/*.002*
		$\alpha = 10^{-4}$	.0001/*.0004*	.0001/*.0002*
4	.15	$\alpha = 10^{-2}$	.010/*.010*	.009/*.009*
		$\alpha = 10^{-3}$	.0009/*.001*	.001/*.001*
		$\alpha = 10^{-4}$	.0001/*.0001*	.0001/*.0001*
	.45	$\alpha = 10^{-2}$	.008/*.011*	.008/*.010*
		$\alpha = 10^{-3}$	.001/*.001*	.0009/*.001*
		$\alpha = 10^{-4}$	.00009/*.0001*	.00007/*.0001*
	.80	$\alpha = 10^{-2}$	.01/*.01*	.009/*.01*
		$\alpha = 10^{-3}$	.001/*.002*	.001/*.001*
		$\alpha = 10^{-4}$	.0001/*.0002*	.0001/*.0002*

is consistent with our simulation results. That is, for some SNPs the χ^2 as obtained when all sibling data are used is up to a factor of 1.85 larger than the χ^2 as obtained when the analysis is limited to siblings with observed genotypes. The χ^2 averaged over the 112 SNPs was $\chi^2 = 2.499$ in the imputed sample, a value larger than the average χ^2 obtained based on the 'observed sample' ($\chi^2 = 2.285$).

Importantly, the results also indicate that the value of test statistic may drop following imputation[4] (i.e., the points below the diagonal in Fig. 4). We conjectured that this drop in value

[4] As an additional check, the analysis of height data was repeated in Merlin (Abecasis et al. 2002), and this analysis produced similar results (results not shown).

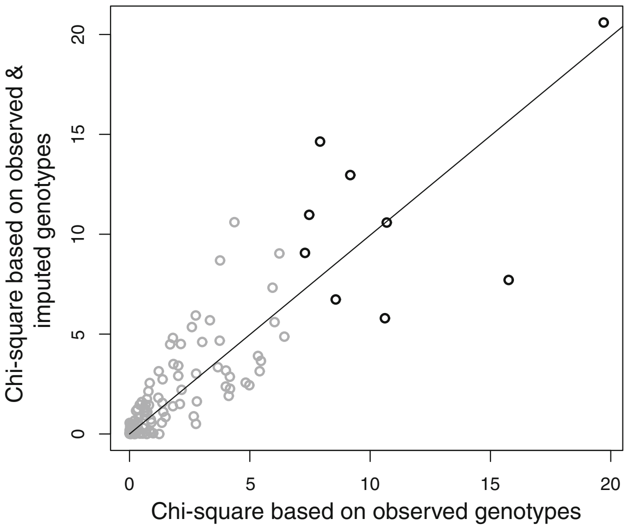

Fig. 4 Chi-square values obtained in the analysis that incorporates 3,500 imputed sibling genotypes along with the 2,410 observed genotypes relative to the chi-square values obtained in the "no imputation analysis". In the latter analysis the sample is limited to the 2,410 observed sibling genotypes. 112 SNPs were tested for association with height. Shown in *black* are the 9 hits at α = .01 based on the observed data. *Points below the diagonal* are due to drop in test statistic following imputation

is due to the small effect sizes given that the 180 SNPs identified explain only about 10 % of the height variance (Lango Allen et al. 2010). While power is increased by the imputation, the actual test statistic may still drop in value, as it remains a single realization of the distribution of the test statistic. This is more likely to occur if the gain in power is relatively small. To test this, we carried out additional simulations.

Additional simulations: explaining height results

Genotypes and phenotypes of a trait with heritability of 80 % (provided that the heritability of height has been estimated at about 80 %, Silventoinen et al. 2003) were simulated for 100 samples consisting of 500 MZ and 500 DZ families with size 4 sibships. The effect sizes of the genetic variants were varied such that they explained .1 %, .5 % and 1 % variance in the phenotype. To mimic the height data we also varied the percent of missingness among the observed parental and sibling genotypes: 50 % genotypes were missing among parent 1 and parent 2 and 25, 60, 90, and 95 % genotypes were missing among sibling 1, sibling 2, sibling 3 and sibling 4, respectively.

Table 6 Increase in χ^2 obtained in a family-based association analysis that includes 2,410 observed and 3,500 imputed sibling genotypes, relative to an association analysis limited to the observed genotypes. The first 4 SNPs are hits at α = .01, the SNP rs1351394 is a Bonferroni significant result

SNP	χ^2 (no imputation analysis)	χ^2 (imputed siblings included)	χ^2 increase
rs1351164	7.467	10.972	1.47
rs724016	9.174	12.967	1.41
rs4282339	7.289	9.063	1.24
rs7759938	7.918	14.640	1.85
rs1351394	19.711	20.599	1.05

In the first step, we imputed the missing sibling genotypes conditional on the observed genotypic data. We then ran the association analyses in each of the three samples: the full information sample, where all siblings ($N = 4{,}000$) had complete phenotype and genotype data, the imputed sample, consisting of siblings with observed ($N = \sim 1{,}600$) and imputed genotypes ($N = \sim 2{,}400$), and the limited sample, where missing genotypes were not imputed ($N = \sim 1{,}600$ genotypes observed). Figure 5 displays the results.

The χ^2 trend as obtained in the three samples was expected to decrease as the genotypic information decreases, with the

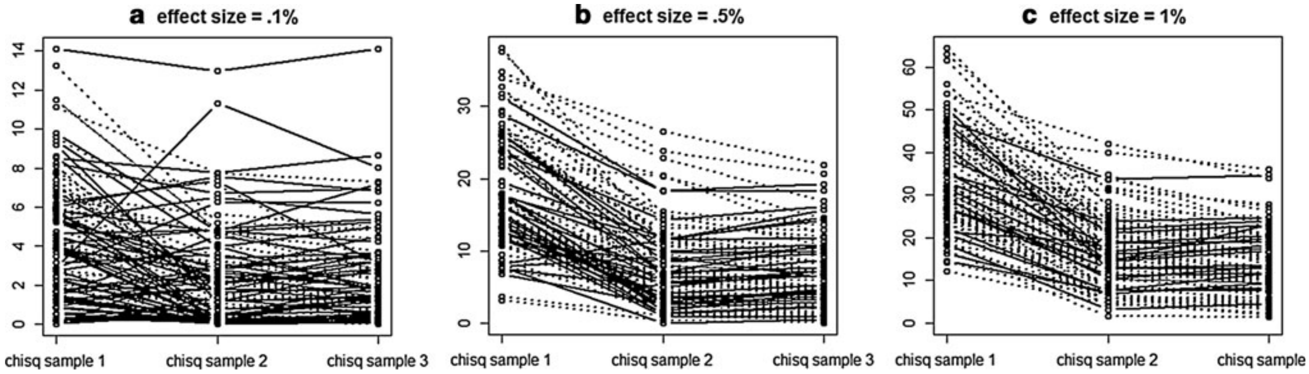

Fig. 5 Chi-square as obtained in three samples: sample 1, consisting of siblings with complete phenotype and genotype data ($N = 4{,}000$), sample 2, consisting of siblings with observed ($N = \sim 1{,}600$) and imputed genotypes ($N = \sim 2{,}400$), and sample 3, where missing genotypes were not imputed ($N = \sim 1{,}600$ observed genotypes). Results are shown for three effect sizes (100 simulated samples). The *dotted lines* show analyses where the chi-square as obtained in the three samples is monotonically decreasing, as expected. The *continuous lines* show results inconsistent with this expectation

imputed sample yielding a χ^2 value that is intermediate between those obtained in the full information setting and in the limited sample. We found that, for the .1 % effect size case, we observed this trend in only 39 % of the analyses, these results are shown as dotted lines in Fig. 5a. However, an increase in the size of the effect was accompanied by an increase in the proportion of results consistent with the expected rank ordering of the χ^2 values; that is, in the .5 % (1 %) effect size case the trend was monotonically decreasing in 67 % (80 %) of the analyses (Fig. 5b and 5c). It follows from these results that the most likely explanation for the drop in test statistic following imputation is the small effect sizes of the 112 SNPs accompanied by large standard errors of the relevant parameter. That the effect sizes are small, in fact too small to be detected in the present sample is evident in the fact that there were only 9 hits based on the observed data—displayed as black points in Fig. 4—at the very liberal alpha of .01.

The analysis of smoking initiation

Results for the association analysis of smoking initiation are given in Table 7.

The results are comparable to those obtained in the previous analysis of height: some SNPs, but not all, showed an increase in the test statistic with the inclusion into analysis of imputed sibling genotypes. Notable is the increase in power obtained at SNP rs3949478, whose association signal approached the significance threshold of $\alpha = 4\text{e-}04$ based on the observed data and reached a *p*-value of 7.27×10^{-06} by including the imputed genotypes. This SNP, located in the ENTPD1 gene, significantly predicts the probability of switching from never-smoking to smoking initiation, conditional on sex and age, after the Bonferroni correction has been applied ($\alpha = .01/20$).

Discussion

Results of the present study suggest the following conclusions and recommendations concerning the use of family-based genotype imputation in genomewide association studies (GWAS). First, we found the mixture model and the dosage approach accommodate equally well the uncertainty of the imputed genotypes. That is, adding imputed sibling genotypes—either by making full use of the distribution of the imputed genotypes or by using genotype dosages—produced unbiased estimates of the parameter of interest. Furthermore, the power of the two

Table 7 Results of 20 tests of genetic association with smoking initiation, ran in the 'complete data' sample ($N = 2{,}458$) and in the sample that includes additionally imputed siblings genotypes ($N = 5{,}981$). Sibling data only were included into analysis. The background covariance matrix was modeled by an AE model. The model was fitted by means of quasi-likelihood and provided Wald-type tests of effects (*t* tests shown in italics), which, for consistency, were converted to χ^2 values

CHR	SNP	No imputation analysis		Imputed siblings added	
		χ^2 (*t* value)	*p*-value	χ^2 (*t* value)	*p*-value
2	rs4608580	.86 (*.93*)	.35	.008 (*.09*)	.92
2	rs10865016	7.18 (*2.68*)	.007	7.61 (*2.76*)	.0057
2	rs787151	9.42 (*3.07*)	.002	11.49 (*3.39*)	.0007
3	rs1599903	.82 (*.91*)	.36	1.06 (*1.03*)	.29
3	rs9824246	.008 (*.09*)	.92	1.21 (*1.10*)	.27
3	rs16860281	7.02 (*−2.65*)	.008	6.20 (*−2.49*)	.01
7	rs6960379	2.49 (*−1.58*)	.11	1.16 (*−1.08*)	.27
7	rs2237781	5.61 (*2.37*)	.01	4.79 (*2.19*)	.02
7	rs4725563	.82 (*−.91*)	.36	.64 (*−.80*)	.41
8	rs4509385	.03 (*−.18*)	.85	.16 (*.41*)	.67
10	rs10999845	1.08 (*1.04*)	.29	.79 (*.89*)	.37
10	rs3949478	12.74 (*−3.57*)	.0004	20.16 (*−4.49*)	7.27e−06
10	rs1856801	.88 (*.94*)	.34	.64 (*.80*)	.42
10	rs7082195	.36 (*.60*)	.54	.13 (*.37*)	.70
11	rs17477949	4.45 (*2.11*)	.03	4.66 (*2.16*)	.03
11	rs12797615	4.92 (*2.22*)	.02	5.95 (*2.44*)	.01
12	rs7313149	2.01 (*−1.42*)	.15	1.82 (*−1.35*)	.17
14	rs8009082	.46 (*−.68*)	.49	.94 (*−.97*)	.32
14	rs8019291	1.04 (*−1.02*)	.30	.92 (*−.96*)	.33
15	rs4774925	2.19 (*1.48*)	.13	1.10 (*1.05*)	.29

approaches was equal across the conditions which were considered. Our findings confirm the results of Visscher and Duffy (2006), who carried out a small scale study of the mixture approach limited to 10 replications. They are also in accordance with the findings of Zheng et al. (2011), who considered the mixture and the dosage approaches in the context of genotype imputation of single nucleotide polymorphism markers (Scheet and Stephens 2006), and found the difference to be small, except given large effects and poor imputation precision. The comparison was performed under an additive genetic model; though, we expect the two approaches would perform equally well also under a non-additive genetic model, as shown by Zheng et al. (2011). All things being equal, the dosage approach is arguably the model of choice in analyzing family data with missing genotypes, as it is computationally more convenient. However, the more demanding mixture approach might prove advantageous in certain circumstances. For instance, this approach could be used to carry out within-family tests of association, allowing one to tackle with stratification (Fulker et al. 1999; Abecasis et al. 2000).

Results of simulations confirmed that the inclusion in an association analysis of imputed sibling genotypes may increase the statistical power. Therefore, for phenotypes for which the siblings resemble each other either weakly (phenotypic correlation $< .4$) or strongly (phenotypic correlation $> .6$) one should consider the inclusion into analysis of imputed genotypes as this approach may increase the power up to a factor of 1.3 relative to the “no imputation analysis”. These gains will be greater if the imputation is informed by observed genotypes in more family members and at more loci—in which case the identical-by-descent information can be exploited to impute siblings with higher accuracy, as demonstrated by Chen and Abecasis (2007) and by Burdick et al. (2006).

Li et al. (2009) noted the advantage of imputation: “(...) imputing genotypes for known relatives of the individuals included in a GWAS of mostly unrelated individuals will always increase power (...) and should be considered whenever phenotyped relatives for the individuals to be genotyped in a scan are available” (page 391). However, the computational effort is not always rewarded by significant gains in power. Specifically, as discussed by Visscher and Duffy (2006), we found the yield of this procedure to low if the phenotypic correlations among the siblings are between about .4 and .6.

As the gains in power also depend on the precision of imputation, the question arises which individuals, if genotyped, would provide maximum information about the missing genotypes in their relatives? The question can possibly be answered by considering the distance from the unconditional H–W genotype probabilities to the probabilities based on the observed genotypes in the relatives. Kinghorn’s genetic probability index (GPI) can be used to express this distance (Kinghorn 1997; see also Percy and Kinghorn 2005), as it equals zero if the imputed probabilities equal the H–W probabilities, and 100 if any genotype probability equals 1. To illustrate this, we used the R library GeneticsPed (Gorjanc and Henderson 2007) to calculate the GPI of the probabilities in Table 1. For instance, in the small example of Table 1, we find that the precision of the imputation is greatest given observed sib AA genotype and observed parent AA genotype (GPI = 86.67), and smallest given sib genotype aa and parent genotype Aa (GPI = 26.69). In contrast, a single observed AA sib confers more information that an Aa sib and an aa parent (GPI 49.33 vs. 41.38). Given that the GPI is approximately related to power, in principle this index provides a means to allocate genotyping resources (Kinghorn 1999). See also Chen and Abecasis (2007) for discussion and illustration of efficient allocation of genotyping resources in multi-locus family based imputation.

Second, we investigated how statistical modeling of the background covariance matrix affected the power to detect a measured (imputed) genetic effect. For low to moderate background correlations, the likelihood ratio test in the linear mixed model appeared to perform correctly when the residual structure was misspecified. Yet, the validity of this conclusion should be considered as confined to the settings of the simulation studies: the analysis was restricted to sibling data, a small effect size of 1 % explained phenotypic variance, heritabilities of .15 and .45. How robust the test is in circumstances different from those considered here (i.e., in larger pedigrees or given larger effect sizes) is subject to further study. Careful specification of the residual structure, however, is required when the trait of interest is highly heritable, as in this circumstance, the misspecification will give more false positives than expected.

Finally, concerning the empirical results we note the following. The imputation will change the distribution of the test statistics under the alternative hypothesis (effect is present), such that the power increases. How much the power increases depends on the background phenotypic correlation among siblings, the number of additional imputed cases, and on the quality of the imputation in terms of the GPI. We note that the actual observed test statistic following imputation need not necessarily be larger than the value of the test statistic observed prior to imputation. As a single realization of the distribution of the test statistic it is likely to be larger if the imputation greatly increases the power. Conversely, if the power benefit is small, that the change in distribution of the test statistic under the alternative is relatively small, the probability is greater of obtaining a smaller value. As the genetic effects are typically hypothesized to be small, in practice, the decision on whether or not family-based imputation should be used as a means to increase power should be informed by prior power calculations and by the consideration of the background correlation.

Acknowledgments Camelia C. Minică and Jacqueline M. Vink are supported by the ERC starting Grant 284167. The statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org) which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003), the Dutch Brain Foundation and the Department of Psychology and Education of the VU University Amsterdam. Data collection and genotyping were funded by the Netherlands Organization for Scientific Research (NWO: MagW/ZonMW grants 904-61-090, 985-10-002, 904-61-193, 480-04-004, 400-05-717, Addiction-31160008 Middelgroot-911-09-032, Spinozapremie 56-464-14192), Center for Medical Systems Biology (CSMB, NWO Genomics), NBIC/BioAssist/RK(2008.024), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL, 184.021.007), the VU University's Institute for Health and Care Research (EMGO+) and Neuroscience Campus Amsterdam (NCA), the European Science Foundation (ESF, EU/QLRT-2001-01254), the European Community's Seventh Framework Program (FP7/2007–2013), ENGAGE (HEALTH-F4-2007-201413); the European Science Council (ERC Advanced, 230374), Rutgers University Cell and DNA Repository (NIMH U24 MH068457-06), the Avera Institute, Sioux Falls, South Dakota (USA), the National Institutes of Health (NIH, R01D0042157-01A), the Genetic Association Information Network (GAIN) of the Foundation for the US National Institutes of Health, the (NIMH, MH081802) and by the Grand Opportunity Grants 1RC2MH089951-01 and 1RC2 MH089995-01 from the NIMH. The authors have no conflict of interest to declare.

References

Abecasis GR, Cardon LR, Cookson WO (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66(1):279–292

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30(1):97–101

Boker S, Neale MC, Maes H, Wilde M, Spiegel M, Brick T, Spies J, Estabrook R, Kenny S, Bates T, Mehta P, Fox J (2011) OpenMx: an open source extended structural equation modeling framework. Psychometrika 76(2):306–317

Boomsma DI, de Geus EJK, Vink JM, Stubbe JH, Distel MA, Hottenga JJ, Posthuma D, van Beijsterveldt TCEM, Hudziak JJ, Bartels M, Willemsen G (2006) Netherlands Twin Register: from twins to twin families. Twin Res Hum Genet 9(6):849–857

Burdick JT, Chen WM, Abecasis GR, Cheung VG (2006) In silico method for inferring genotypes in pedigrees. Nat Genet 38(9): 1002–1004

Chen WM, Abecasis GR (2007) Family-based association tests for genomewide association scans. Am J Hum Genet 81(5):913–926

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Prentice Hall, Harlow

Fulker D, Cherny S, Sham P, Hewitt J (1999) Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64(1):259–267

Gorjanc G, Henderson DA, with code contributions by Kinghorn B and Percy A (2007) GeneticsPed: Pedigree and genetic relationship functions. R package version 1.20.0. http://rgenetics.org

Kinghorn BP (1997) An index of information content for genotype probabilities derived from segregation analysis. Genetics 145(2):479–483

Kinghorn BP (1999) Use of segregation analysis to reduce genotyping costs. J Anim Breed Genet 116(3):175–180

Laird NM, Lange C (2011) The fundamentals of modern statistical genetics. Springer Verlag, New York

Lango Allen H, Estrada K, Lettre G, Berndt S, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Ferreira T, Wood AR et al (2010) Hundreds of variants influence human height and cluster within genomic loci and biological pathways. Nature 467(7317):832–838

Li Y, Willer C, Sanna S, Abecasis GR (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406

Mather K, Jinks JL (1977) Introduction to biometrical genetics. Cambridge University Press, Cambridge

Percy A, Kinghorn BP (2005) A genotype probability index for multiple alleles and haplotypes. J Anim Breed Genet 122(6): 387–392

Pinheiro J, Bates D, DebRoy S, Sarkar D, the R Development Core Team (2012) nlme: linear and nonlinear mixed effects models. R package version 3.1–104

R development Core Team (2005) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78(4):629–644

Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, Dunkel L, De Lange M, Harris JR, Hjelmborg JV, Luciano M, Martin NG, Mortensen J, Nisticò L, Pedersen NL, Skytthe A, Spector TD, Stazi MA, Willemsen G, Kaprio J (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res 6(5):399–408

Van der Sluis S, Dolan CV, Neale CM, Posthuma D (2008) Power calculations using exact data simulation: a useful tool for genetic study designs. Behav Genet 38(2):202–211

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York

Vink JM et al (2009) Genome-wide association study of smoking initiation and current smoking. Am J Hum Genet 84(3):367–379

Visscher PM, Duffy DL (2006) The value of relatives with phenotypes but missing genotypes in association studies for quantitative traits. Genet Epidemiol 30(1):30–36

Visscher PM, Benyamin B, White J (2004) The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. Twin Res 7(6):670–674

Visscher PM, Macgregor S, Benyamin B, Zhu G, Gordon S, Medland S, Hill WG, Hottenga JJ, Willemsen G, Boomsma DI, Liu YZ, Deng HW, Montgomery GW, Martin NG (2007) Genome partitioning of genetic variation for height from 11,214 sibling pairs. Am J Hum Genet 81(5):1104–1110

Visscher PM, Andrew TA, Nyholt DR (2008) Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. Eur J Hum Genet 16(3): 387–390

Willemsen G, de Geus EJC, Bartels M, van Beijsterveldt TCEM, Brooks AI, van Burk GFE, Fugman DA, Hoekstra C, Hottenga JJ, Kluft K, Meijer P, Montgomery GW, Rizzu P, Sondervan D, Smit AB, Spijker S, Suchiman HED, Tischfield JA, Lehner T, Slagboom PE, Boomsma DI (2010) The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. Twin Res Hum Genet 13(3):231–245

Zheng J, Yun L, Abecasis GR, Scheet P (2011) A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genet Epidemiol 35(2):102–111